Labels:

Level 1: Abstraction, Nature, Concepts & Ideas

Labels:

Level 1: People, Places, Nature, Work & Occupations, Leisure & Pastimes

MULTI-LABEL CLASSIFICATION OF ARTWORKS INTO THEIR REPRESENTED ICON(S)

Author:
Ludovica Schaerf
Institution:
Amsterdam University College
Supervisor:
Giovanni Colavizza

Outline

RESEARCH QUESTIONS

inspiration: ILSVRC
motivation: Annotation speed-up

METHODS

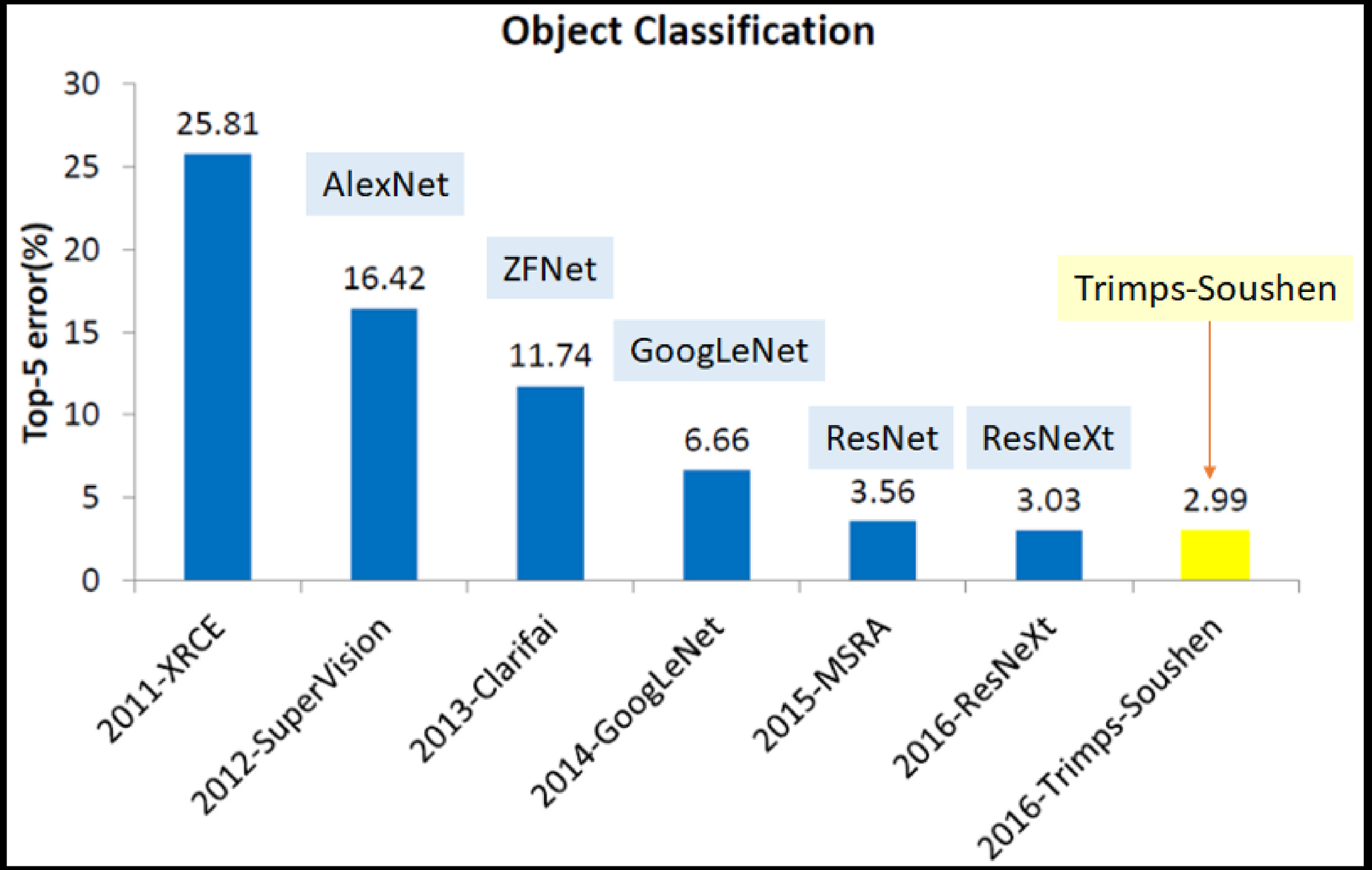
transfer learning
multi-label methods
dataset

RESULTS

what model and what pre-trained were best?
what classes were best/worst classified?
what periods/movements were best/worst classifed?

IMPLICATIONS AND LIMITATIONS

practical utility
state of deep learning algorithms
future research



ILSVRC (2010-2017): ImageNet Large Scale Visual Recognition Challenges

Deep Residual Neural Networks (2015): 'outperform human ability'.

How well can AI's recognise the content of artworks?

WHAT MODEL BEST RECOGNISES THE CONTENT OF ARTWORKS?

which pre-trained model?
which classification architecture?

WHAT ELEMENTS ARE MOST PROBLEMATIC FOR THE MODEL?

which icons/classes?
which periods?
which movements?

WHAT ABOUT ABSTRACT ART?

how is the performance compared to other classes?
how well can the model recognise their representational or conceptual content?

Motivation

- Automate/Speed up the annotation process
- Reflect on the ability of machines to recognise the content of artworks

Transfer Learning

FROM IMAGES TO PAINTINGS

- source data: ImageNet
- pre-trained: VGG, InceptionV3, ResNet
- fine-tuning

Andy Warhol

Mick Jagger

1980

Arts, Inc. / Artists Right Society (ARS),

EXPLORE

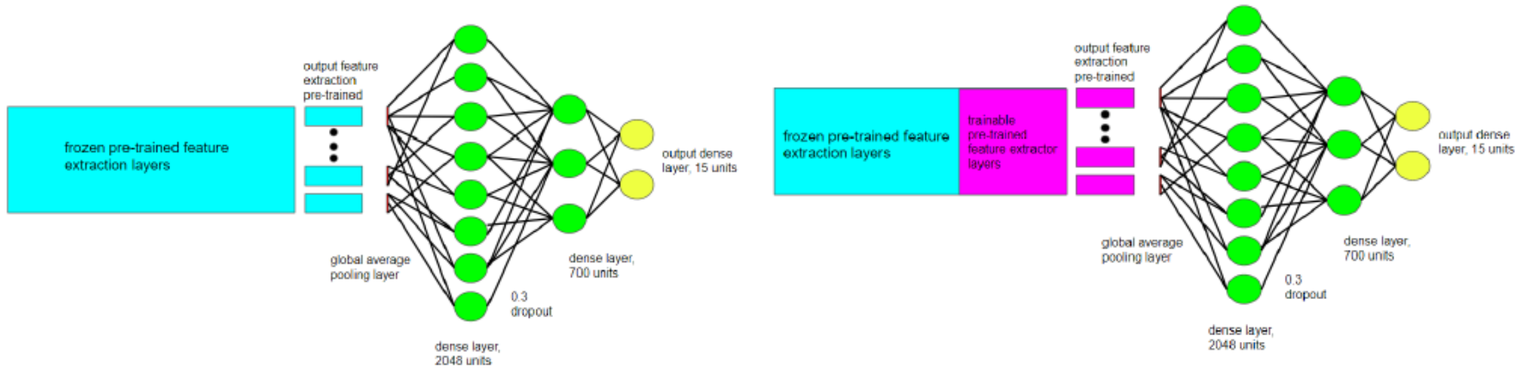
emotions, concepts and ideas (16,874)
formal qualities (12,931)
colour (876)
photographic (5,089)
history (5,803)
arts (1,795)
exhibition: 'Andy Warhol', Artcurial, Centre d'Art Plastique Contemporain, Paris, 1980 (1)
leisure and pastimes (7,669)
art and craft (2,355)
exhibition (1,376)
music and entertainment (2,170)
music, pop (60)
objects (23,378)
fine art and design, named works (6,078)
Warhol, Andy, print (38)
reading, writing, printed matter (5,151)
poster (878)
people (35,142)
adults (22,736)
man (10,497)
body (4,942)
head / face (2,549)
named individuals (12,533)
Jagger, Mick (25)
portraits (4,470)
individuals: male (1,965)
places (41,542)
cities, towns, villages (non-UK) (13,291)
Paris, Avenue Matignon 9, Artcurial Centre d'art Plastqiue Contemporain (1)
countries and continents (17,557)
France (3,518)
society (34,906)
lifestyle and culture (10,260)
advertising (537)
symbols and personifications (7,212)
inscriptions (6,720)
name of artist (199)
printed text (1,145)
work and occupations (14,164)
arts and entertainment (6,586)
singer (209)

Multi-label Classification

FROM MULTI-CLASS TO MULTI-LABEL

Naive Multi-label Classification:

- Binary Cross-Entropy Loss
- Sigmoid



(a) Off-the-shelf model architecture

(b) Fine-tuned model architecture

Figure 3.1: A visualisation of the model architectures both in the case of fine-tuning and not. As visible in the picture, only the feature extractor layers of the pre-trained are used, while the classification is added by us. The classification is made up of a global average pooling layer, two hidden fully connected layers and an output layer.

Multi-label Classification

FROM MULTI-CLASS TO MULTI-LABEL

CNN-RNN (Wang et al, 2016):

- CNN: image-to-label embedding
- RNN: label-to-label embedding
  - k-nearest neighbour
  - beam search

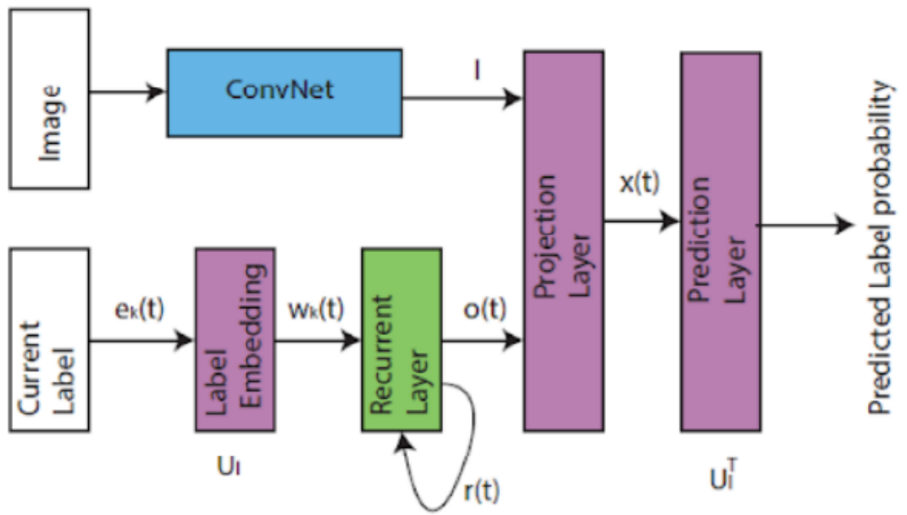


Figure 3.6: An illustration of the architecture of CNN-RNN [30]

Tate Dataset

- 25'000 images
- British art 1900-2020
- Sketches, photographs and paintings
- by Tate and National Galleries of Scotland

- 15 Classes:
- People, Objects, Places, Nature, Abstraction, ..., Emotions and Ideas, Work and Occupations

Results

WHICH PRE-TRAINED AND WHICH CLASSIFICATION METHOD PERFORM BEST?

Model	O-A	O-F1	O-P	O-R	P-A	P-F1	P-P	P-R
Baseline	0.83	0.27	0.37	0.25	-	-	-	-
VGG16 no tuning	0.85	0.51	0.60	0.48	0.71	0.32	0.39	0.27
VGG16 class weights	**0.86**	0.50	**0.62**	0.48	0.71	0.32	0.37	0.34
VGG16 fine tuning	0.85	**0.52**	0.58	**0.51**	0.71	0.34	0.41	0.31
Inception V3 no tuning	0.82	0.37	0.50	0.34	0.73	0.34	0.40	0.29
Inception V3 class weights	0.83	0.40	0.51	0.35	0.74	0.33	0.42	0.27
Inception V3 fine tuning	0.85	0.49	**0.62**	0.44	0.72	0.32	0.37	0.29
ResNet50 no tuning	0.81	0.42	0.47	0.42	0.71	0.34	0.36	0.33
ResNet50 class weights	0.83	0.42	0.52	0.39	0.72	0.34	0.38	0.31
ResNet50 fine tuning	0.84	0.47	0.57	0.43	0.71	0.31	0.35	0.28
CNN-RNN	0.82	**0.52**	0.57	**0.51**	**0.83**	**0.54**	**0.57**	**0.51**

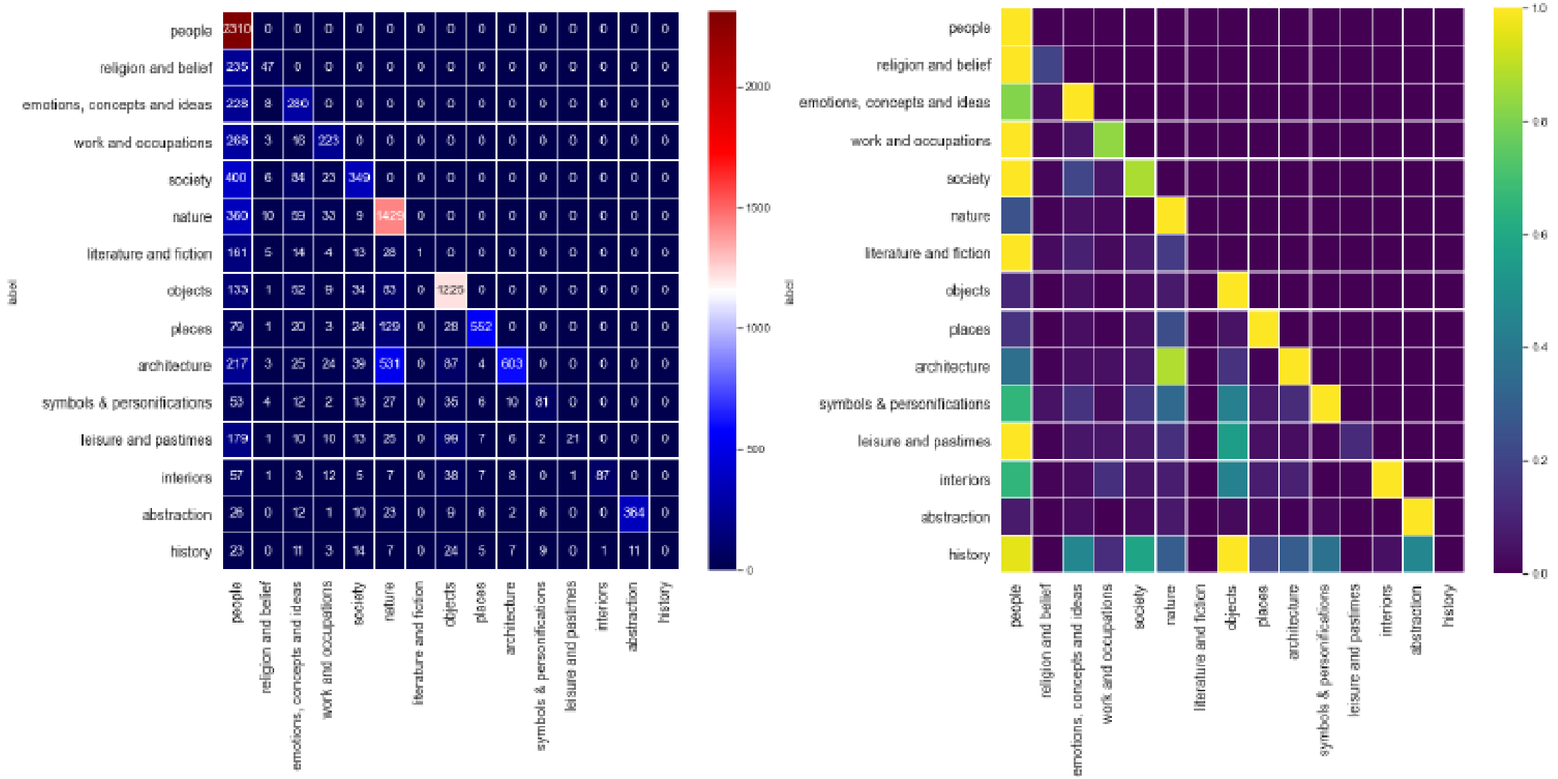
Table 5.1: Results Table. The results are organised according to the pre-trained used and the training method. *No tuning* refers to the models trained with the convolutional layers of the pre-trained frozen, *fine tuning* is the model in which 1/2 of the convolutional layers are re-trained and *class weights* is the model, with no tuning, that is trained with a weight associated to each class. Finally, the CNN-RNN model uses VGG as pre-trained.

Example Predictions

predicted as: ['<start>', 'places', 'abstraction', '<end>', '<end>'],

ual class: ['<start>', 'objects', 'abstraction', 'emotions, concepts and ideas', '<

What classes are most problematic?



(a) Absolute Confusion Matrix Heat-map of CNN-RNN

(b) Relative Confusion Matrix Heat-map of CNN-RNN

Figure 5.2: A heat-map visualisation of a confusion matrix of the predictions of CNN-RNN. Being a multi-label classification, the confusion matrix is computed by adding 1 if the prediction of label i of the image j is 1 and the ground-truth is also 1. When the ground-truth is 1, but it has not been predicted, to all of the mis-predictions are added $1/n_j$ where n_j is the number of mis-predictions on the image j. The normalisation is done row-wise, by dividing each entry in a row by the maximum of that row. NB: The classes are not in any specific order.

What periods are most problematic?

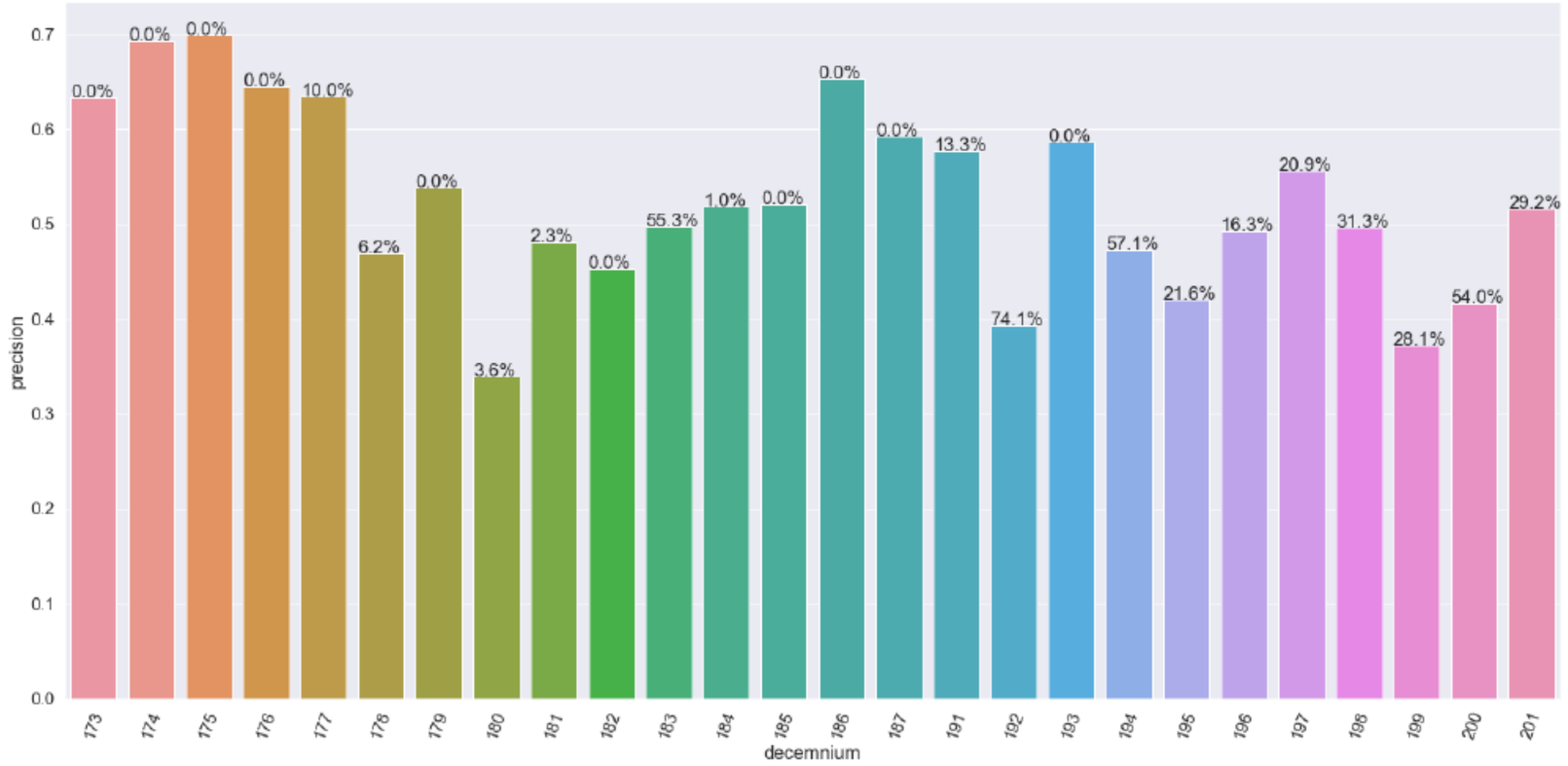


Figure 5.4: A bar plot with confidence intervals of the precision per year of CNN-RNN. The percentage on each bar represents the coverage of the movement data for that decennial.

What movements are most problematic?

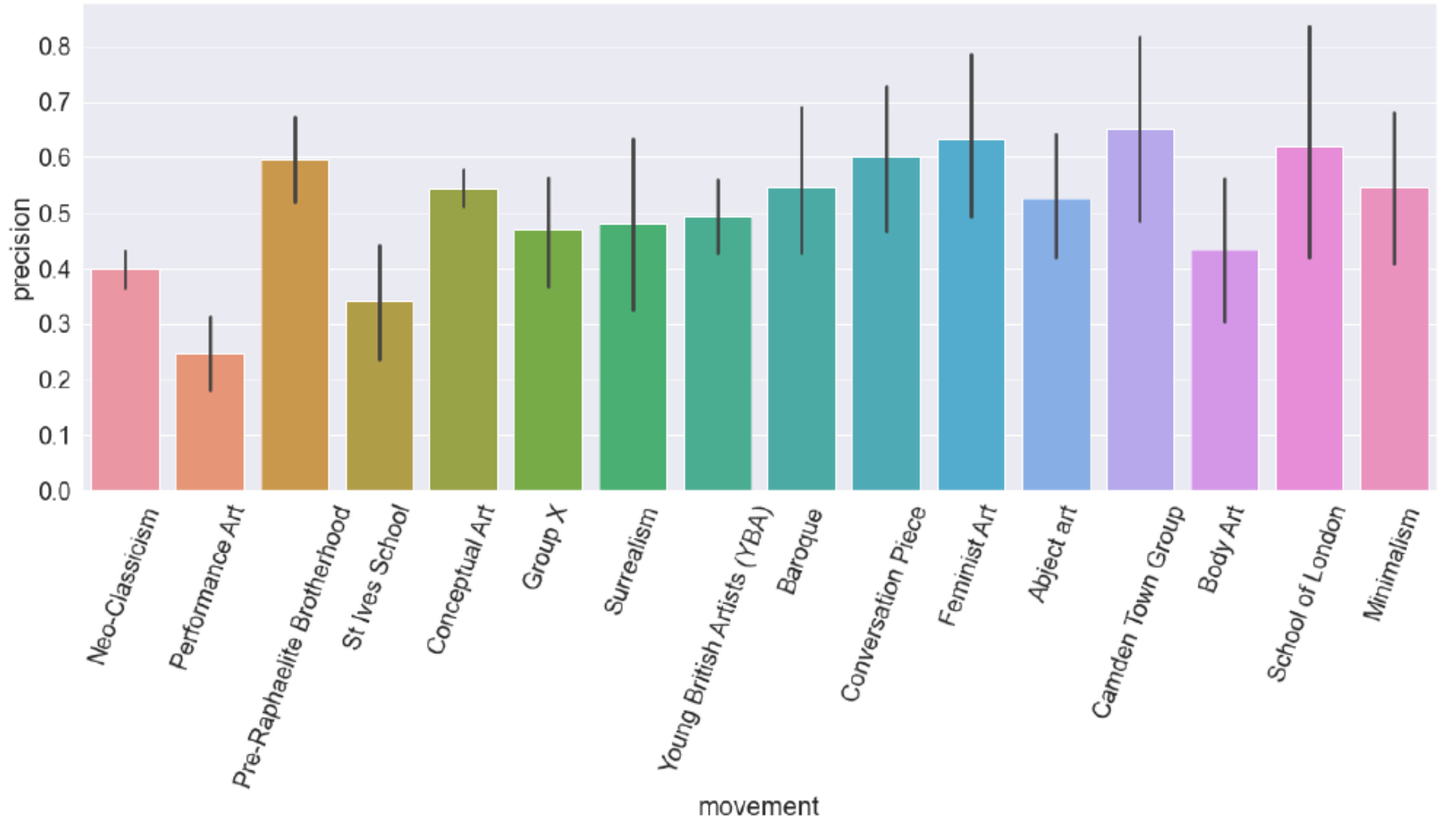
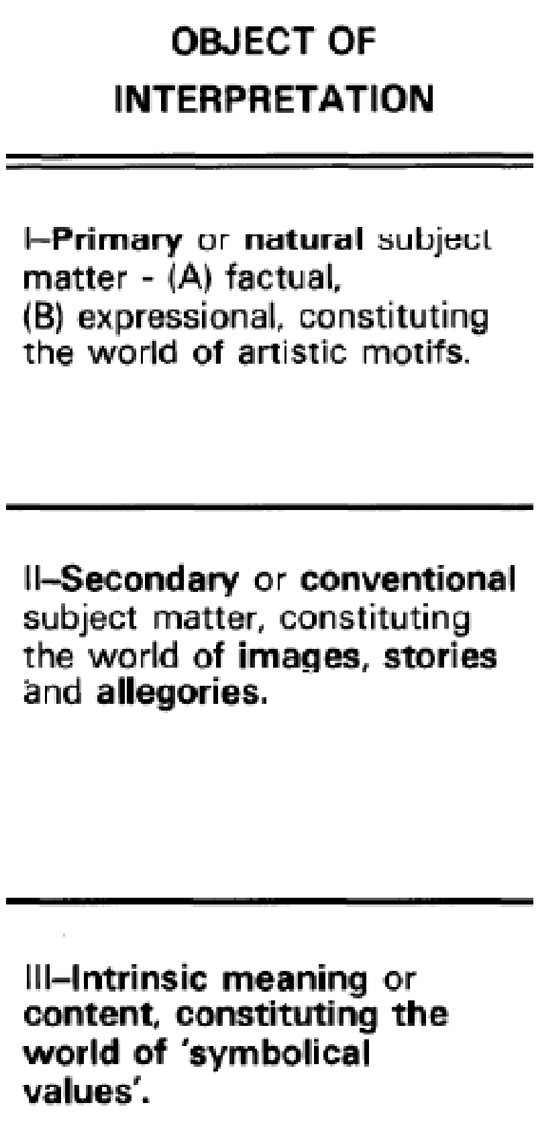


Figure 5.5: A bar plot with confidence intervals of the precision per movement of CNN-RNN. Only the movements that appear more than 10 times in the test set are kept.



Implications

THE MODEL IS NOT READY TO BE PUT IN PRODUCTION

Suggestion-Revision mechanism

PRIMARY, AND SECONDARY, SUBJECT MATTER

The model can extrapolate the manner in which objects, and concepts, are expressed with shapes

FOR THE FUTURE

Data scarcity
Class imbalance
Hierarchical methods